Local Flexibility Markets: Future-proofed

Sebastian Blake, Maddie Brooks • 3 July 2023[1]

Introduction

The benefits of using demand side flexibility to smartly manage distribution network congestion are well established. Much progress has been made in unearthing the system value of local flexibility in the GB market with all DSOs now procuring flexibility services on a commercial business-as-usual basis, as an alternative to traditional grid upgrades. There are a number of market mechanisms utilised by DSOs to manage network congestion at present, and increasing conflicts between these local flexibility market mechanisms and with national markets. Blunt approaches which were correct to adopt when fostering a market from scratch present scaling challenges, therefore, now is the right time to assess whether this cocktail of mechanisms are delivering the right signals to manage the complex demands of our future electricity network.

To some commentators a smart demand flexibility is a fait accompli, i.e. given substantial benefit is capturable from responding to crude market signals in place today surely all new EVs and heat pumps will arrive 'ready smart' and the system benefits will naturally materialise. This is not true. Building automated flexibility requires developing complex architecture (e.g. for handling interoperability & cyber security, understanding the trade off between opposing market signals, processing speeds to match the real time nature of power balancing). This activity is capital intensive and capability must be designed against uncertain future market requirements, the incentive on flexibility operators is effectively to deliver sooner through cruder functionality. To be clear, advanced & extensive demand flexibility is a necessity, not a choice, for Net Zero and without the right market signals 'sufficiently smart' flexibility will not appear organically.

System value

Any system, from motorways to telecommunications, must have sufficient capacity to manage peak load or users will experience a disruption to the service. Given electricity demand is forecasted to grow by ~50% from today by 2035 (FES EC.02) left uncoordinated this demand could, at worst, increase peak load; driving lower system efficiency and reliability. However, this demand has the potential to deliver far greater benefits than simply not contributing to the national network peak; for example responding to local network peaks, making best of renewable energy, insulating against sudden outages, stabilising voltage and more.

[1] Contacts: sebastian.blake@octoenergy.com or madelaine.brooks@octoenergy.com

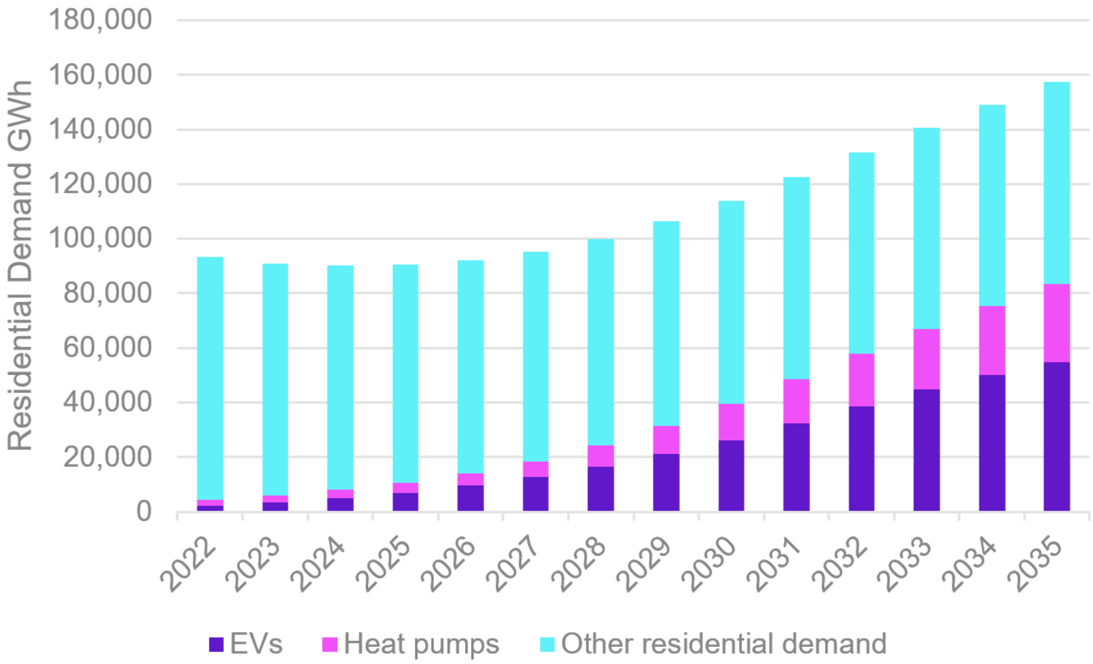

Figure 1: Data from NGESO's FES 2020 Consumer Transformation scenario - showing the contribution of EV and Heat Pump demand to total residential demand.

The demand growth will happen at the grid edge, e.g. within homes and businesses, largely driven by the electrification of heating and transport and it is now well proven these technologies have significant flexibility in their operation. This means they are highly responsive to price signals (especially if controlled via automation); so with the clear markets we can unlock a cheaper, cleaner grid at a rapid pace.

In this paper we will focus on the use case of using resources at the grid edge to respond to conditions of distribution network (termed 'local flexibility'), however also considering how this interplays with other valuable use cases (e.g. at the national system level). Local flexibility brings value primarily through permanently reducing the need for greater network capacity, especially as it is deployable on a faster timescale than traditional infrastructure. It is a dynamic and low regret solution which can be implemented quickly to give network operators a varied toolset for managing investment (e.g. often buying optionality to understand where network upgrades are truly required). Studies like Imperial's "Net-zero GB electricity: cost-optimal generation and storage mix[2]" have concluded that capital expenditure associated with this avoided asset upgrade is around £1bn per annum for GB.

A whole system approach

When drawing conclusions about the value of local flexibility it is essential to consider the demands of the whole electricity system. For example, there is a tension between optimising locally for maximum grid efficiency (e.g. a flat demand) and optimising based on national signals like coordinating demand with highest instances of renewable output

[2] Page 17, Distribution Capex costs only, Aunedi, M. et al. (2021) Net-zero GB electricity: cost-optimal generation and storage mix. doi: 10.25561/88966

(see diagram below). This potential conflict between maximising consumption of low cost generation whilst staying within network limits has not been considered in upfront market design (e.g. siloed approaches to baselining), and only recently have discussions emerged about how best to coordinate. This issue has been largely avoidable to date given the broad alignment of national and local signals (i.e. evening peak avoidance). As large volumes of flexible load arrives it destroys this assumption, therefore good market design should acknowledge the competing use cases and find the most efficient solution to it.

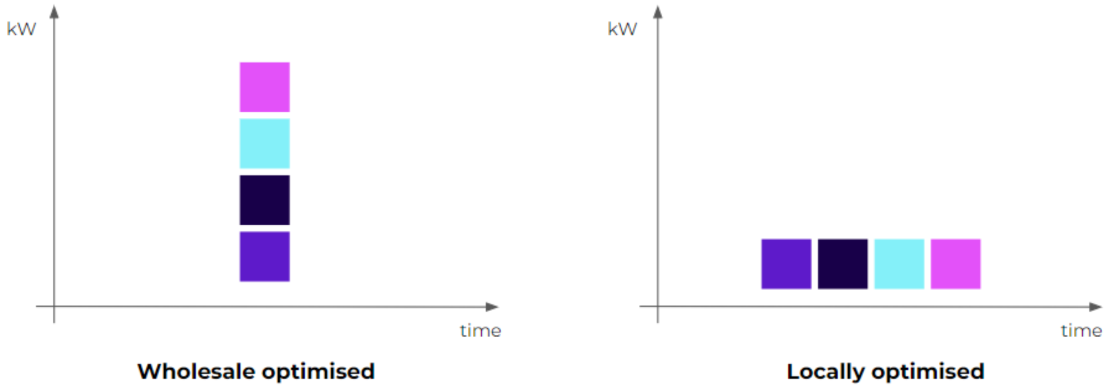

Figure 2: Representative diagram to show the different outcomes between optimising based on the single wholesale price vs optimising to maximise grid efficiency.

Asset flexibility

The greatest opportunity in local flexibility lies with heat pumps and EVs. They are both highly flexible assets with a large degree of 'zero cost' optionality in their consumption profile. For example, an EV driver simply cares that their car is full and ready to drive in the morning but it is of zero impact whether this charging happens at, say, 12am or 3am. Here we are simply shifting the timing of the consumption, the total amount of energy consumed is unaffected by acting flexibly. As long as this scheduling is administered in a low cost way (i.e. via algorithm rather than manual instruction) this flexibility has a low activation cost.

Figure 3: Graph which shows the price responsiveness of heat pumps demand on Octopus Energy's Cosy tariff

However, a low activation cost should not encourage complacency and a belief that a small price signal, which might already be present in the current wholesale market, is all that is needed. As discussed, the systems required to enable the necessary automation have a high upfront cost, without a clear market signal this flexibility will not appear organically.

These assets may also have some degree of interruptible flexibility as well, for example accepting a slightly unfilled EV or slightly colder home in return for payment. However, this flexibility is likely to be small in comparison and has a high activation cost. Therefore, its role is likely to be far more limited.

Assets with a low activation cost can adapt constantly to price signals, continuously re-optimising in real time to lower costs. This presents a difficulty for market designs which rely heavily upon a baselining. How do you define the normal 'expected' behaviour of an asset which is constantly changing its position based upon market conditions? Many DSO's have adopted a pragmatic approach of a fixed baseline, i.e. assuming the behaviour of asset class

Local flexibility market mechanisms

There are three core market design options for local flexibility we will explore (other options can be conceived but the following have the most merit). In fact all three are deployed in the GB market today, however sometimes creating conflicting signals.

1. **Congestion pricing** - a volumetric signal which adds to the price of electricity at times of high congestion to disincentivise consumption (or incentivise generation).
2. **Flexible connections** - grid users accepting non-firm access to the network in return for some incentive (eg a cheaper or faster connection).

3. **Contracted flexibility service** - a service agreement where grid users agree to shift from expected behaviour, in exchange for a payment.

State of the market

1. Static DUoS

A static congestion pricing signal has been long present for distribution network users via the volumetric Red, Amber and Green DUoS charges. This provides a daily peak time signal for both generation and demand, which varies by region and voltage level. The static DUoS price signals are updated annually.

Static DUOS provides a generally helpful signal to reduce peak demand however it only broadly represents actual congestion on the network given the level of temporal and locational aggregation. One set of prices applies (depending on connection level) across the distribution network zone (or GSP Group), meaning within each zone highly congested parts of the network are incentivised equally as parts with much spare capacity. In addition, there is no seasonal variation in the price signal so the same incentive applies in summer as in winter, when in summer the evening peak demand might be well below the maximum capacity. Given the price signal is spread out over the whole year and across the network zone it is low in magnitude; e.g. compared to if it were concentrated to only occurrences of actual congestion. For this reason, static DUoS signals are rarely passed through directly to domestic consumers (e.g. via time of use tariffs or flexibility products to request turn down).

2. Flexible Connections (Active Network Management)

Flexible connections have proved a valuable tool for connecting renewable energy (and occasionally large points of demand) without the grid upgrades which would be required to support a firm connection. In GB this is also known as Active Network Management (ANM) and essentially is where a distribution connected user agrees to have its output curtailed by the DSO when an export constraint appears and without compensation for this event.

Flexible connections have allowed many generators to connect far earlier to the network by avoiding the required grid infrastructure for a firm connection. ANM is mainly dispatched on a 'last in, first out' methodology where the latest user to connect with ANM is the first prioritised for curtailment (and so on down the list). When this happens there is a missed opportunity to instead find a use for this otherwise curtailed energy, e.g. heat pumps could be coordinated to use this energy for water heating.

It is important to note here that any instance of renewable energy curtailment is not proof of system inefficiency; as it would be highly inefficient to build enough redundant storage / grid infrastructure to eliminate all instances of curtailment. However, an efficient system should first look to see if any useful purpose for this energy can be found elsewhere before curtailment occurs.

There have been important positive developments in this area via the Access SCR. Whereas previously curtailing renewable generation was effectively a free option for the DSO, the updated incentive mechanism penalises its excessive use and rewards curtailment efficiency. This creates the incentives where it was missing previous, which leads to new market mechanisms as UKPN are now tendering for demand turn up services as an alternative to curtailing generation (80 MW in their latest tender), and other DSO will surely follow.

Ofgem has also introduced principles to better manage the uncertainty of a flexible connection from a provider's perspective (e.g. DSOs must provide a limit on how often the connection can be curtailed) which increases the financeability to projects with flexible connections. ANM data will also be made public under RIIO-ED2 both measures allowing market participants better visibility of their connection risk.

One growing issue for ANM sites is that there is confusion as to how the flexible connection interacts with other flexibility markets, like ESO balancing and the Capacity Market. For example, when new services like DC have been launched by ESO it has excluded ANM sites from participating - and although this position has evolved there is still a frustrating lack of coordination between national markets and local flexibility market mechanisms which increases complexity and ultimately reduces liquidity in all impacted markets.

3. Flexibility tenders

All GB DSOs now tender for flexibility services on a regular basis as an alternative to traditional grid upgrades. The table below shows volumes procured across pre-fault (Sustain & Secure), post-fault (Dynamic & Restore) & reactive services. Octopus Energy has won tenders to deliver flexibility with each DSO so is well placed to review the experiences of these market arrangements.

DSO Tenders (Industry Total)	**Sustain (MW)**	**Secure (MW)**	**Dynamic (MW)**	**Restore (MW)**	**Reactive Power (MVAr) (if applicable)**
	Peak Capacity (MW)	Peak Capacity (MW)	Peak Capacity (MW)	Peak Capacity (MW)	Peak Capacity (MVAr)
Contracted for 2018	0	24	34	59	0
Contracted for 2019	0	10	121	125	0
Contracted for 2020	2	105	556	502	0
Tendered for 2020	14	493	771	778	7
Contracted for 2021/22	**28**	**375**	**926**	**538**	**0**
Tendered for 2021/22	**57**	**840**	**1584**	**1149**	**5**
Contracted for 2022/23*	**37**	**192**	**643**	**220**	**0**
Tendered for 2022/23*	**141**	**802**	**1289**	**961**	**11**

*** Contracted/Tendered to date, more expected over the remainder of 2022**

Figure 4: ENA Open Networks Table summarising the capacity tendered and contracted under each DSO flexibility product category.

While DSO procured local flexibility has now progressed to regular commercial tenders it is still an evolving marketspace. Some of the issues which are now deserving of attention, such as a lack of standardisation across DSOs, are highlighted in Ofgem's Future of Distributed Flexibility Consultation (along with potential fixes). There are certain no regrets options which should be adopted, however, before committing to any of the more 'big bang' solutions to coordination we should ask whether alternative market designs could be more appropriate in a future system, or could sit alongside DSO flexibility tenders.

Some of the issues with flexibility tenders are more inherent to tendered services and not easily fixed through different coordination structures like shared asset registries. For example, while the contracting burden on participants can be lightened through adoption of a common set of terms across DSOs, a tendered service would always require some form of contract and therefore creates a barrier in comparison to an explicit exposure to a price signal.

One of the greatest issues with flexibility services is baselining. Baselining is a deep and complex field which deserves greater focus than we have space for here (and all too often ignored completely in discussion). Essentially, a service which offers a payment to adjust output must determine a counterfactual case where the instruction wasn't sent (the baseline) in order to establish how much delivery has occurred. They key point to understand is baselining demand is a difficult exercise, which gets dramatically harder over smaller sample sizes. Local flexibility is by definition interested in small catchments where baselining can be almost impossible, e.g. an individual household will have a highly erratic demand profile and it is only through aggregating to 100s of households that portfolio effects start to smooth this out. Given the difficulty of accurately describing a range of asset behaviour via simple formula baselining is highly susceptible to gaming. Baselining is a necessary aspect of demand flexibility to define how much it has occurred, but designs which place an explicit requirement on the market operator to calculate this (contracted services) are more greatly impacted by their weaknesses than ones which do not (price signals).

Another key issue is herding. Procuring turn down services over the evening period only works so long as the peak occurs in the evening time. What do we do when 50% of demand on the low voltage network is from EVs and heat pumps which are highly responsive to market signals. We can derive a contracted service to move this volume out of 'peak' time but given the volume this will simply create a new peak later on, creating an endless game of 'chase the peak'. Flexibility tender currently have no incentive for flexibility providers to introduce diversity, such that it is valuable to the overall system, and at worse disincentivises it given providers can earn payments for resolving constraints.

It is easy to envisage how dishonest actors could manipulate tendered services to extract maximum compensation. For example, deliberately coordinating asset consumption to cause a constraint, in order to earn a payment for then resolving. If different assets were used to cause and alleviate the constraint then monitoring for this behaviour would be difficult. While this example may seem particularly egregious, other gaming strategies may be more subtle but ultimately more damaging if more easily deployed at scale.

Another consideration for a tendered service approach is how this structure rewards certain players. Adopters of low carbon technologies like heat pumps and EVs (usually the most able to pay) effectively cause the network constraints but are then rewarded for resolving them. The cost of these services is then socialised across all users meaning the poorest consumers end up paying more to fund flexibility services and grid upgrades which they didn't cause or directly benefit from.

Improving the status quo

There are different approaches to improve the status quo for local flexibility. Ofgem's Future of Distributed Flexibility consultation lays out some options for improving the structural framework around flexibility tenders, ranging from evolutionary to big bang system coordination.

We must also consider reform to network charging and flexible connections at the same time however, or risk a bad overall design. All these signals stack together and poor coordination can explicitly block market access for certain players, results in sub-optimal operational optimisation decisions and create a confusing landscape for flexibility providers and investors.

We can split the options for market design reform into three buckets.

1. Evolutionary changes

These are low regret options which should be advanced regardless of the ultimate market design pursued. These are well covered elsewhere (such as the Open Networks project) and items like standardisation across DSOs (for baselining rules, contract terms, APIs etc) undoubtedly should have been delivered already.

Such tactical changes are helpful but all too often used as justification for delaying more meaningful progress.

2. Central system coordination

More ambitious options put forward in Ofgem's Future of Distributed Flexibility Consultation envisage a central asset registry or even a single flexibility platform where upward and downward response can be traded.

There is much to be attracted to by these concepts. The flexibility marketplace is incredibly fragmented with a vast number of price signals and services to incentivise essentially the same activity (see the flexibility markets diagram).

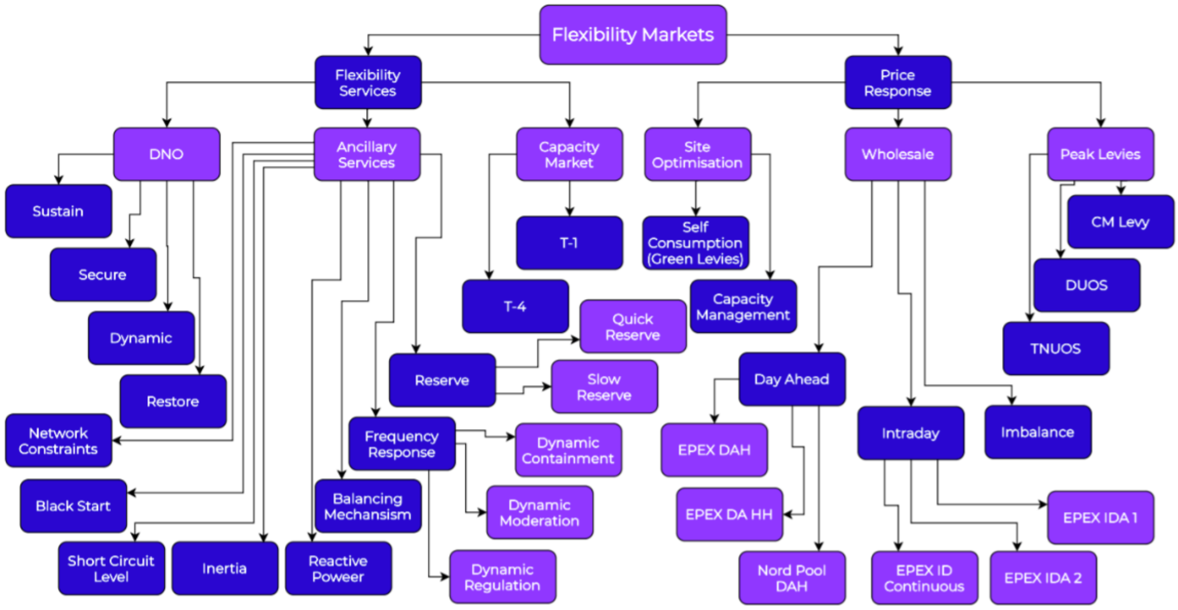

Figure 5: Key flexibility markets in GB (note other bespoke services exist but not shown).

Efficient market participation is hampered by the poor mapping of aggregated units between markets. A DSR provider will have different units in DSO services, the Capacity Market and the Balancing Mechanism. With arbitrary aggregation rules (e.g. component reallocation in the Capacity Market) preventing the provider keeping units consistent. A centralised (or multiple but coordinated) asset registry should hold asset information independent of the market provider to facilitate interoperability (i.e. MPAN database).

We believe such grander attempts at better market coordination deserve greater investigation; as a participant having several platforms to sell your flexibility on concurrently is far from ideal. However, to mitigate the lengthy design and delivery risk we should first look to expand the role of existing systems where possible. With facilitation of greater access to the BM could this system not provide the fundamental control and dispatch for most flexibility markets? For example, DNOs dispatching ANM via the BM would force far greater coordination between ESO and DSO than years of primacy workshops have achieved.

3. Alternative market designs

Whilst we recommend rapid implementation of low regrets options and rapid scoping of larger coordination solutions; we will give particular space to consideration of alternative market structures for local flexibility, given the lack of attention in current discourse. The alternative we believe deserves most merit is dynamic congestion pricing; we will also briefly examine nodal pricing within the distribution network and dynamic operating envelopes.

Dynamic DUoS

If the static DUOS signal were adapted to be more reflective of real system congestion a far sharper price signal could be passed through. This would disincentivise consumption, or incentivise generation, at times of congestion through an administratively set price signal, much like current DUOS, but unlock greater benefits through being highly responsive to actual system conditions.

Unlike static DUOS, where the signal is watered down across the year and network zone, the dynamic price signal would only come into effect when real congestion is detected. This would greatly sharpen the price signal, and greatly disincentivise consumption when demand is highest and encourage consumption when the grid is underutilised. The gradient of the price signal, see below, is a trade off between sharpness and facilitating forecasting, but this could change over time as forecasting became more sophisticated. The overall price level would be set such that it reflects the cost of upgrading the constrained infrastructure.

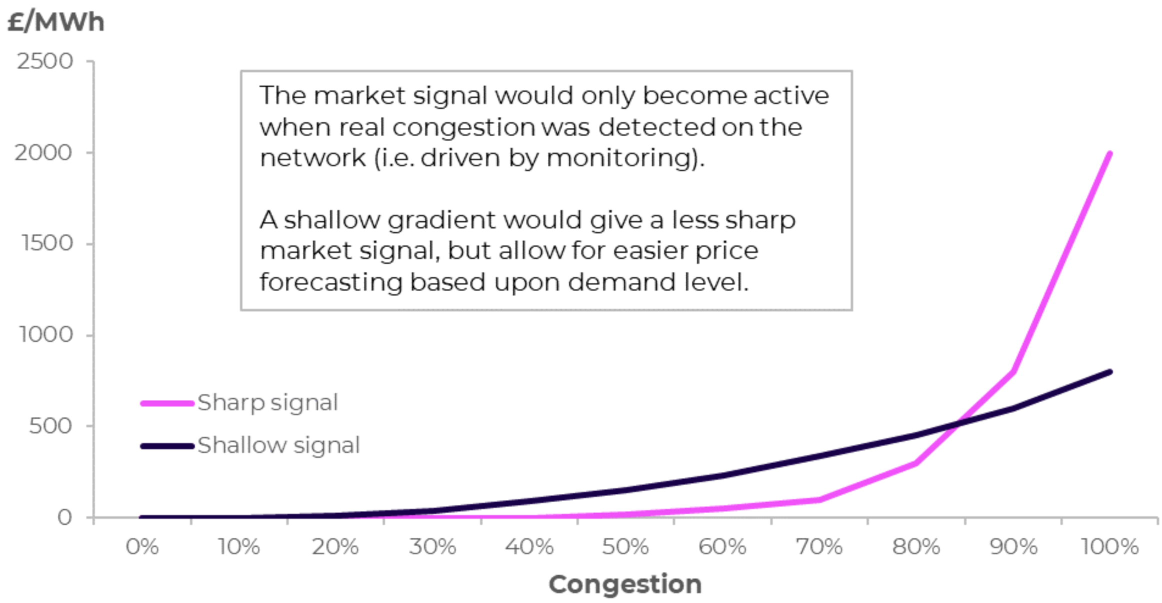

Figure 6: A representative chart showing how congestion at a network bottleneck could drive pricing.

Driving the signal with real time system monitoring would deliver the greatest benefit but imputed congestion could drive the calculation initially if easier to implement. Once in place however, the congestion risk is passed on to market participants who must train forecast models based on public congestion data to optimise assets in real time. Making the transition to this regime sooner would allow market participants to build up this forecasting capability while demand is still driven by highly forecastable patterns, indeed for some regions with consistent demand shapes the dynamic DUoS signal would be very similar to the static DUoS signal initially before evolving.

This would bring an array of benefits:

- Herding - 'chasing the peak' would cease to become an issue as the congestion price is reflexive in nature, i.e. would naturally update to reflect changing system conditions. A truly reflexive signal would need to be set by actual real time congestion, and only be determined ex-post (like the System Imbalance Price).
- Gaming - the design is more resilient to anti-competitive behaviour as inducing congestion wouldn't create any opportunity to earn revenue from the DSO. Equally, baseline manipulation would cease as an issue given this model doesn't depend on regulated baselining formulae for market operation.

- Market driven - this approach would take pressure off network operators to forecast and design services to resolve constraints. While there would be an initial ramp of capability from market participants, in the long run it would prove the most dynamic mechanism to the rapidly changing nature of local grid conditions.
- Fairness - EVs and heat pumps would directly face the congestion costs they cause, and responding to them would lower the costs faced by all users. Or one more targeted approach to mitigate this risk could be to only enter the dynamic pricing regime once a household has purchased a heat pump / EV of sufficient size (i.e. only for connections above a certain rating).

In theory, a signal could replace almost all the local flexibility signals currently present (network charging, flexible connections and flexibility tenders) however in reality it is most likely to be explored in conjunction with other signals already present. In principle it can deliver a far more coordinated signal to actors than the current mixed regime does.

There are drawbacks of a price-based flexibility market it is important to note:

- Fairness - congested parts of the network would experience higher prices in general creating price differentials across the network. These challenges are explored in the LMP debate, in general we believe the political challenge of price differentials can be swallowed if the benefit overall to consumers is clear.
- Price certainty - flexibility operators will have much greater risk in comparison to contracted flexibility. This would present a barrier where capital investment is being considered in a specific location (e.g. purchasing a battery). Capital investment which isn't locationally specific (e.g. developing EV control strategies) is less impacted.
- Set up complexity - a congestion model of the whole network would be needed if the mechanism were to be implemented everywhere with a mapping to price signal. This becomes a complex exercise particularly where the network is meshed, therefore certain locations are likely to be better candidates initially.
- Forecastability - flexibility operators are not best placed (at least initially) to produce forecasting models (needed to predict congestion prices). So transferring this risk to flexibility operators would only outperform a centralised approach over time as skills and information processing was built up.

Other mechanisms

Other market design options for local flexibility are conceivable, we will examine a couple below.

One idea could be to apply nodal wholesale pricing within the distribution network, so pricing at nodes would reflect the distribution constraints. There are a number of disadvantages to this approach but it can be most easily ruled out through market power considerations, i.e. there would be many instances of price making power being concentrated with one or two generators (and there is much evidence of how traders respond to this situation). Market power is a non-trivial issue for any competitive power market where network constraints exists, e.g. the current wholesale market allows competition between actors when physically this competition doesn't exist but simply shifts the issue into the Balancing Mechanism. Allowing generators to exhibit control over wholesale pricing deep within the distribution network would be highly difficult to regulate against.

Another idea, termed dynamic operating envelopes, is a form of flexible connection, essentially a more advanced form of ANM where the DSO issues a dynamic capacity allowance to each site (rather than a simple 'last in, first out'). AusGrid have effectively deployed this solution to optimally manage domestic solar exports, however naturally there is much hesitance about how such a tool could apply to demand. For good reasons, curtailing generation creates lost revenue (which could be compensated) but not being able to heat a home or charge a car can deliver unacceptable levels of disruption for electricity users. This solution has similar drawbacks to ANM, there is no market mechanism to optimally dispatch assets in merit order and centralising this decision making in the DSO prevents the market from finding efficient solutions (like using otherwise curtailed solar for heating water in the middle of the day).

Conclusion

Demand growth will be predominantly driven by the uptake of EVs and heat pumps, and these technologies have proven significant capability to provide flexibility. Given the doubling to tripling in demand at the grid edge that's expected in the coming years we cannot depend entirely on network upgrades to manage this growth. In the worst case consumers will be unable to decarbonise through purchase on an EV or heat pump due to grid capacity.

Whilst a number of different mechanisms are currently being used to defer the need to upgrade network infrastructure including congestion pricing, flexible connections and DSO contracted flexibility services, there are a number of emerging issues and complexities that this complex web of mechanisms is producing. Now is the time to explore if there are alternative market mechanisms which are more capable of dealing with this increasingly complex system, where historic tools which rely on diversity to predict maximum demand will no longer be sufficient as the uptake of smart devices responding in real time to price signals becomes the norm.

Price signals must become more reflective of actual system conditions in order to elicit a useful response from these assets, that is helpful at both a local and system level. Dynamic DUoS is most resilient to gaming, truly reflexive, and capable of replacing almost all the local flexibility mechanisms currently operational, delivering a far more straightforward mechanism which can more easily be integrated with national markets to mitigate conflicts. Importantly, it has the ability to solve inefficiencies across products which the current siloed approach to flexibility markets creates.

We must use the upcoming Ofgem DUoS Significant Code Review process to test a number of different and innovative ways to manage local constraints to find the optimum market design as soon as possible. The exponential growth of heat pumps and EVs is upon us so now is the time to respond and plan for it; and Kodak-like thinking will allow the less desirable impacts of cruder signals to manifest at scale, ultimately draining industry effort to fix easily forecastable problems. The prize of local flexibility is seismic but only achievable through advanced market design. Few topics have similar potential to unlock deep decarbonisation across sectors and hence are so deserving of our attention.